Implementing an Open-Source Document Preservation System at NASA GES-DISC

AGU Fall 2013 Meeting
San Francisco, CA December 9-13 2013

mo.khayat@nasa.gov
http://disc.gsfc.nasa.gov

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

A system for the preservation of documentation for legacy ESDIS missions at GES-DISC

Mo Khayat[1,2], Steve Kempler[1], Barbara Deshong[1,3], Irina Gerasimov[1,3], James Johnson[1,3], Ed Esfandiari[1,3], Michael Berganski[1,3]

[1]NASA Goddard Space Flight Center, [2]Wyle Science, Technology & Engineering [3]ADNET Systems Inc.

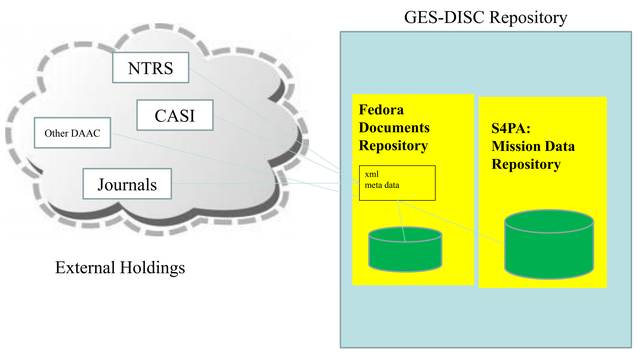
Why the need for preservation?

Many Earth Observing System (EOS) missions have either reached the end of their active life or are nearing it. Preservation of data and artifacts from these missions is critical to the long-term studies of our planet's climate, and to aid future generation's ability to understand climatic changes. Data from these legacy missions provide valuable time comparisons of environmental conditions without which our climatology analysis would be incomplete. A significant hurdle in making these comparisons is the ability to understand and interpret data from the older missions when the experts familiar with them have moved on or are long gone. NASA has recognized the importance of preserving the Earth Science data with the issuance of the "NASA Earth Science Data Preservation Content Specification" (423-SPEC-001) for the Earth Science Data and Information System (ESDIS) supported data centers. The GES-DISC is one of the ESDIS data centers that has been actively pursuing this preservation directive over the last few years. To this end we have setup the GES-DISC Repository System, which is capable of long-term archive of documentation artifacts and other associated digital content. For considerations of cost and web services flexibility, we designed this system based on Fedora Commons, an open-source repository management software. We will also present our concept of operations for user access and detail how our implementation takes into account limitations on document access, based on considerations of proprietary or sensitive information. The first mission to make use of the GES-DISC Repository System is the High Resolution Dynamics Limb Sounder (HIRDLS) on the Aura spacecraft. In this poster we provide status of this implementation with lessons learned and our plans to incorporate other legacy and ongoing missions.

Motivation

NASA remote sensing data are a national resource with great scientific value that needs to be preserved and shared for future scientific research, by generations to come. Maintaining and ensuring their use is essential as we learn new ways to utilize these data in science research and applications.

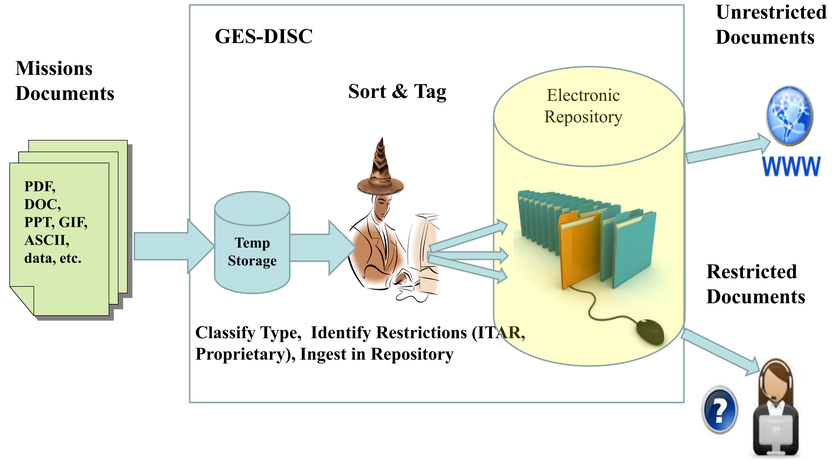
No preservation = Loss of Future Long Term Climate Records



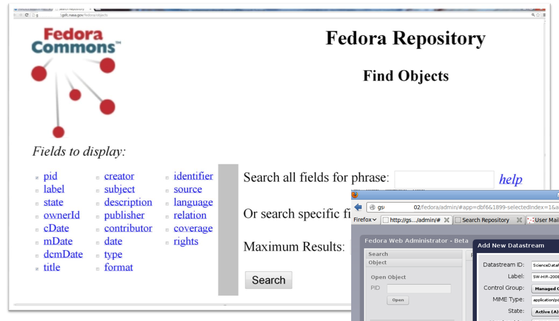
Documents in the GES-DISC repository are archived and classified according to the 423-SPEC-001 into 8 categories listed in the table below for each mission.

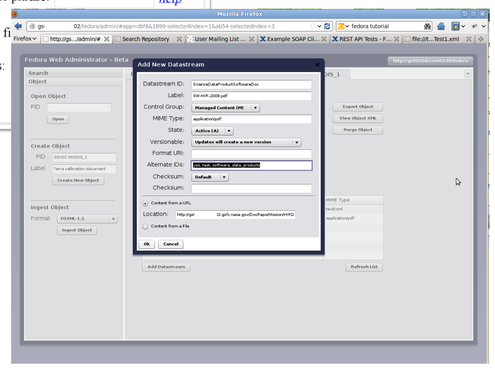
1. Category	2. Content Item	3. Definition/Description
Preflight/Pre-Operations Calibration	Instrument Description	Documentation of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.).
	Preflight/Pre-operational Calibration Data	Numeric (digital data) files of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.).
Science Data Products	Raw Data and Derived Products	Raw data are data values at full resolution as directly measured by a spaceborne, airborne or *in situ* instrument. Derived products are higher level products (level 1b through 4) where calibration and geo-location transformations have been applied to generate sensor units, and/or algorithms have been applied to generate gridded geophysical parameters.
	Metadata	Information about data to facilitate discovery, search, access, understanding and usage associated with each of the data products.
Science Data Product Documentation	Product Team	Names of key science team leads and product team members (development, help desk and operations), roles, performing organization, contact information, sponsoring agencies or organizations and comments about the products.
	Product Requirements	Requirements and designs for each science data product, either explicitly or by reference to the requirements/design documents. Product requirements and designs should include content, format, latency, accuracy and quality.
	Processing and Algorithm Version History	For all products held in the archive, documentation of processing history and production version history, indicating which versions were used when, why different versions came about, and what the improvements were from version to version. For all products held in the archive, the versions of source code used to produce the products should be available at the archive.
	Product Generation Algorithm	Detailed discussion of processing algorithms, outputs, error budgets and limitations. Processing algorithms and their theoretical (scientific and mathematical) basis, including complete description of any sampling or mapping algorithm used in creation of the product , geo-location, radiometric calibration, geophysical parameters, sampling or mapping algorithms used in creation of the product, algorithm software documentation, & high-level data flow diagrams. Description of how the algorithm is numerically implemented.
	Product Quality	Description of the impact to product quality due to issues with computationally intensive operations (e.g., large matrix inversions, truncation and rounding). Documentation of product quality assessment (methods used, assessment summaries for each version of the datasets). Description of embedded data at the granule level including quality flags, product data uncertainty fields, data issues logs, etc. Relevant test reports, reviews, and appraisals.
	Product Application	Useful references to published articles about the use of the data and user feedback received by the science and instrument teams about the products. Includes reports of any peculiarities or notable features observed in the products.
Mission Data Calibration	Calibration Method	The methods used for instrument/sensor radiometric and geometric calibration while in operation (e.g., in orbit). The source code used in applying the calibration algorithms. Documentation of in-line changes to calibration or to instrument or platform operations or conditions that occur throughout the mission.
	Calibration Data	Instrument and platform engineering data collected during operations (e.g., on orbit), including platform and instrument environment, events and maneuvers; attitude and ephemeris; aircraft position; acquisition logs that record data gaps; calibration look-up tables; and any significant external event data that may have impacted the observations.
Science Data Product Software	Science data product generation software and software documentation	Source code used to generate products at all levels in the science data processing system. Software release notes, including references to versions of operating systems, compilers, commercial software libraries used in the code. Versions of science data product software should be archived for each major product release. A major product release is characterized by the appearance of peer reviewed publications where reported results are based on the product version.
Science Data Product Algorithm Inputs	Ancillary data and documentation	Complete information on any ancillary data or other data sets used in generation or calibration of the data set or derived product, either explicitly in data descriptions or by reference to appropriate publications. Ancillary data should be stored with the products unless it is available from another permanent archive facility.
Science Data Product Validation	Datasets and documentation	Accuracy of products, as measured by validation testing, and compared to accuracy requirements. Description of validation process, including identification of validation data sets, measurement protocols, data collection, analysis and accuracy reporting.
Science Data Software Tools	Software and documentation	Product access (reader) tools. Software source code that would facilitate use of the calibration data, ancillary data and the data products at all levels. Includes software source code useful for creating programs that will read and display the calibration data, ancillary data and product data and metadata values. Commercial tools should be identified with appropriate references.

NASA Earth Science Data Preservation Content Specification (423-SPEC-001)



Internal interface used to ingest and archive records into the repository. We have also developed a command line batch mode ingest for multi-document upload into the repository system.





External users access the publicly available documents by visiting the mission specific documentation page for that instrument. The Fedora repository system is at the backend and makes access to the linked documents possible.
http://disc.sci.gsfc.nasa.gov/Aura/additional/documentation/hirdls-preservation-documents

GES DISC Preservation Implementation Status

1. Identify documentation
   GES DISC Science Support staff identified specific information needed per mission in Data Preservation Mission List by working closely with original mission teams to sort out documents for preservation.
2. Specify and implement preservation environment
   Local archive based on open-source system Fedora Commons.
   Implementation is complete for HIRDLS datasets . Accessible to public this month.
   Other missions UARS, MLS, TRMM, AIRS etc. coming soon.
   Exploring NASA Technical Reports Server (NTRS) and NASA Aeronautics and Space Database (NA&SD) as repository for restricted documents
3. Retrieve documentation
   Public documents accessed by users on mission portal pages.
4. Implement retrieval and distribution services
   1. *Access for internal GES-DISC users*
   2. *External Access via WWW for unrestricted documents*
   3. *External Access for restricted documents (ITAR) via User Services contact*
   4. *External Access for restricted documents via authentication (NA&SD)*
   5. *Iterate with other DAACs/community*

Lessons Learned, Challenges, and Future Plans

- Heritage missions require extensive preparatory work to identify and classify documents for preservation
- Restricted vs. Unrestricted documentation to account for ITAR or Proprietary documents requires special handling and accommodation in repository.
- Still investigating viability of NASA infrastructures like NTRS and NA&SD and rules relating to DAA.
- Incorporate DOI metadata into repository.
- Level of service to provide external users if NTRS or NA&SD not a viable option:
- No distribution of restricted documents to external users?
- Distribute only via our User Services on case-by-case basis and only to known domestic users not subject to export control?
- Setup a registration service that complies with export control with authenticated access.
- Do we preserve field campaign data? ACOS data? Model data? MEaSUREs data? Validation datasets?